Retrieval Augmented Generation (RAG)

Large Language Models (LLMs) are being used widely in current Generative AI systems. Unfortunately, LLMs demonstrate significant capabilities but face challenges such as hallucination, outdated knowledge, and non-transparent, untraceable reasoning processes.

Retrieval Augmented Generation (RAG) has emerged as a promising solution by incorporating knowledge from external databases. This enhances the accuracy and credibility of the models, particularly for knowledge-intensive tasks, and allows for continuous knowledge updates and integration of domain-specific information. RAG synergistically merges LLMs’ intrinsic knowledge with the vast, dynamic repositories of external databases. This talk gives an overview of the structure of RAG systems and includes a demo of their capabilities.

Madhu Chinnambeti is an SVP and Senior Data Scientist at SupportVectors. In his current role, Madhu advises the companies, aspiring engineers, and entrepreneurs on ML, AI, and Gen AI technology stack. Madhu has over 28 years of experience in Computer Science and Engineering and he is currently working on his PhD dissertation in the area of Graph Neural Networks (GNNs) and deep learning at Boise State University. Madhu is currently working on research and publications that advance GNNs under Cybersecurity and fraud detection.

His passion also includes tech education in the evolving fields like Generative AI. Madhu is a volunteer advisory board member of Disability:In New Jersey affiliate to help individuals with disabilities to get jobs.